Checkpoint #4: Machine Learning

Casey Grage & Renee Zha

DataBricks Link for #1

1. Given data from years 'A' through 'E,' can we predict how the rate of complaints that result in disciplinary action against the officer will change over time for years 'F' through 'J'? Does this model get better or worse with added features (in addition to year): officer gender, officer race, complainant race, and complainant gender?

I want to see if there is cross-reference between the features outlined above. Can we predict who (of the officers) will get disciplined based on ALL complainant and officer demographics? When we base the model on only complainant identity group, does it make the model better or worse? What about just the officer identity group?

To do this, we compared black and white officers and complainants. We examined all combinations of black, white, male, and female officers/complainants and took the linear regression of each model for the given years. We excluded datasets that were too small to be accurate. We defined "too small" to mean < 200 cases. As a result, we have no linear regressions for female officers and the years for both training and testing the model vary slightly. Interestingly, there were not enough white female complainants against black male officers to extract a linear regression.

Commands 4, 5, 6, and 7 reflect how we gathered the combinations of officer/complainant demographics. Then, we imported the CSVs from DataBricks into Excel and simply used "=SLOPE(<y-value cells, x-value cells>)" to get the linear regressions for the trained and tested data, named "Training Slope" and "Testing Slope," respectively, in the table below. We took the difference between these two slopes to determine how similar/accurate the training model was to the training model.

Officer Race	Officer Gender	Complainant Race	Complainant Gender	Years Used to Train	Training Slope	Years Used to Test	Testing Slope	Difference between slopes
Black				2006-2011	-0.8417075	2012-2015	-0.2297619	0.6119456
White				2006-2011	-0.0634228	2012-2016	-0.3629848	0.299562
		Black		2006-2011	-0.4288327	2012-2016	-0.2514708	0.1773619
		White		2006-2011	-0.341427	2012-2015	-3.6451199	3.3036929
		Black	M	2006-2011	-0.4048764	2012-2015	-0.2086057	0.1962707
		White	M	2006-2011	-0.6295023	2012-2015	-3.3464325	2.7169302

		Black	F	2006-2011	-0.4605658	2012-2015	-0.2131993	0.2473665
		White	F	2006-2011	0.73055228	2012-2015	0.83939082	0.10883854
Black	M	Black	M	2006-2011	-0.8417075	2012-2015	-0.2297619	0.6119456
Black	M	White	M	2006-2011	-2.2251629	2012-2015	-6.7620039	4.536841
Black	M	Black	F	2006-2011	-0.2487561	2012-2015	0.01204519	0.26080129
Black	M	White	F	*not enough data — no single year had more than 132 cases*				
White	M	Black	M	2006-2011	-0.1822687	2012-2015	-0.2000361	0.0177674
White	M	White	M	2006-2011	0.09617566	2012-2015	-1.3780537	1.47422936
White	M	Black	F	2006-2011	0.06927193	2012-2015	0.16332768	0.09405575
White	M	White	F	2007-2011	0.50475525	2012-2015	1.31011053	0.80535528
Avg. Change in Slope with Only Officer Race:						0.4557538		
Avg. Change in Slope with Only Complainant Race:						1.7405274		
Avg. Change in Slope with Either Officer or Complainant Race:						1.0981406		
Avg. Change in Slope with Complainant Race and Gender:						0.817351485		
Avg. Change in Slope with Officer Race and Complainant Race and Gender:						0.97512446		

There is no difference in the accuracy of models when training sets vary based on demographic data. This is because there aren't consistent trends for any of the combos. This means that rates of complaints successfully resulting in disciplinary action, while they do vary with complainant and officer demographics, have not changed with any significant trend in the last 10+ years. Overall, I think this linear regression approach is a good method for analyzing chronological data.

DataBricks Link for #2 & 3

2. Given an officer, and the nature of the complaint (TRR details), can we predict the subject identities? What are the most important features?

I created 3 random forest classifiers: one that only predicted subject_gender, one that only predicted subject_race, and one that predicted subject_gender / subject_race combinations.

Here are the following sorted feature importances for each random forest classifier:

Solely predicting subject_gender:

```
Features sorted by their importance:
[(0.4523, Column<officer_gender>),
 (0.189, Column<officer_in_uniform>),
 (0.1837, Column<weapon_discharged>),
 (0.0769, Column<officer_age>),
 (0.0453, Column<lon>),
 (0.0246, Column<officer_race>),
 (0.0141, Column<officer_assigned_beat>),
 (0.0079, Column<lat>),
 (0.0033, Column<lighting_condition>),
 (0.0014, Column<officer_rank>),
 (0.0013, Column<officer_on_duty>)]
```

Solely predicting subject_race:

```
Features sorted by their importance:
[(0.5113, Column<lat>),
 (0.252, Column<lon>),
 (0.0785, Column<officer_on_duty>),
 (0.0606, Column<officer_age>),
 (0.0504, Column<officer_race>),
 (0.0361, Column<lighting_condition>),
 (0.0078, Column<officer_in_uniform>),
 (0.0023, Column<officer_rank>),
 (0.0007, Column<officer_gender>),
 (0.0002, Column<officer_assigned_beat>),
 (0.0001, Column<weapon_discharged>)]
```

Predicting subject_race/subject_gender combinations:

```
Features sorted by their importance:
[(0.478, Column<lat>),
 (0.2182, Column<lon>),
 (0.1391, Column<officer_gender>),
 (0.0632, Column<officer_age>),
 (0.0386, Column<officer_on_duty>),
 (0.0293, Column<officer_in_uniform>),
 (0.017, Column<lighting_condition>),
 (0.0109, Column<officer_race>),
 (0.0051, Column<weapon_discharged>),
 (0.0006, Column<officer_rank>),
 (0.0, Column<officer_assigned_beat>)]
```

In descending order, the three most important features for predicting subject gender are officer_gender, officer_in_uniform, and weapon_discharged. It seems male subjects are more likely to have TRRs with male officers, and females with females. Also, if a weapon was discharged, the subject is more likely to be a male. My theory with the officer_in_uniform feature importance, is that officers are very often in uniform and the subjects are very often male. So the importance of this feature was overinflated.

In descending order, the four most important features for predicting subject race are location (lat, lon), officer_on_duty, officer_age, and officer_race. Location makes sense because Chicago is so segregated. Whether the officer is on duty or not is also interesting because it shows how many emergency calls there were. Also officer_race is a predictor of subject race. This may be because officers in one area may be more likely to be the same race as the civilians in that area (as per segregation mentioned above). It may also be that white officers are more likely to target black subjects. Officer_age is interesting. Perhaps certain age groups are more likely to have racial biases.

In descending order, the three most important features for predicting subject gender/race combinations are location (lat, lon), officer_gender, and officer_age. For the aggregate subject identity, location and officer demographics are most likely predictors. Location was the number one predictor of subject_race and officer_gender was the number one predictor of subject_gender, so those make sense and have already been discussed. What's interesting is officer race did not make top 3 when predicting subject race and gender combined. It's possible certain age groups are more likely to have racial and gender biases than are race groups to have combined racial and gender biases.

3. Given a subject, and the nature of the complaint (TRR details), can we predict the officer identities? What are the most important features?

Solely predicting officer_gender:

```
Features sorted by their importance:
[(0.5188, 'subject_gender'),
 (0.2699, 'officer_in_uniform'),
 (0.0808, 'officer_assigned_beat'),
 (0.0713, 'lon'),
 (0.033, 'lat'),
 (0.0145, 'subject_race'),
 (0.0047, 'lighting_condition'),
 (0.0034, 'weapon_discharged'),
 (0.0028, 'officer_rank'),
 (0.0007, 'officer_on_duty')]
```

Solely predicting officer_race:

```
Features sorted by their importance:
[(0.4159, 'lon'),
 (0.3472, 'lat'),
 (0.1298, 'subject_race'),
 (0.0447, 'officer_rank'),
 (0.0249, 'officer_on_duty'),
 (0.0175, 'subject_gender'),
 (0.0164, 'lighting_condition'),
 (0.0024, 'officer_in_uniform'),
 (0.0007, 'weapon_discharged'),
 (0.0005, 'officer_assigned_beat')]
```

Predicting subject_race/subject_gender combinations:

```
Features sorted by their importance:
[(0.3772, 'lon'),
 (0.3193, 'lat'),
 (0.1436, 'subject_gender'),
 (0.079, 'subject_race'),
 (0.033, 'officer_rank'),
 (0.0297, 'officer_in_uniform'),
 (0.0116, 'officer_on_duty'),
 (0.0059, 'lighting_condition'),
 (0.0007, 'officer_assigned_beat'),
 (0.0, 'weapon_discharged')]
```

In descending order, the three most important features for predicting officer gender are subject_gender, officer_in_uniform, and officer_assigned_beat. It's really interesting that officer_rank isn't very important at all in determining gender! The genders correlate because again TRRs are more likely to involve officers/subjects of the same gender. Also, it seems whether the officer is in uniform or in their assigned beat correlates with gender. I think male officers are more likely to get called for / take emergency calls when they're not on duty.

In descending order, the three most important features for predicting officer race are location (lat, lon), subject_race, and officer_rank. Again, Chicago is highly segregated so it's possible officers are probably more likely to be the race of their assigned beat (or they're own neighborhood if they took an emergency call). It's also possible certain officers have racial biases against subjects. Officer rank is interesting. Perhaps there is racial bias in the Chicago Police force or men of color are less likely to have higher education and therefore less likely to be promoted in CPD.

In descending order, the three most important features for predicting officer gender/race combinations are location (lat, lon), subject_gender, and subject_race. These are the three features I would expect to be most important given we know location is correlated with race, races are correlated, and genders are correlated.